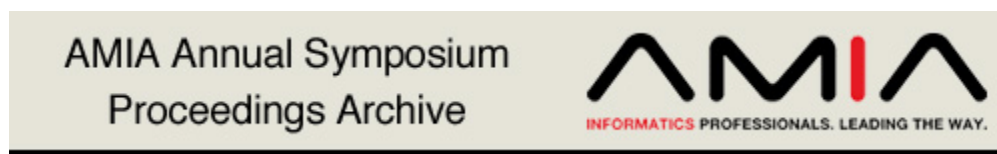


As a library, NLM provides access to scientific literature. Inclusion in an NLM database does not imply endorsement of, or agreement with, the contents by NLM or the National Institutes of Health.

Learn more: [PMC Disclaimer](#) | [PMC Copyright Notice](#)



[AMIA Annu Symp Proc](#). 2018 Dec 5;2018:232–241.

## FAIRness and Usability for Open-access Omics Data Systems

[Daniel C Berrios](#)<sup>1,4</sup>, [Afshin Beheshti](#)<sup>2,4</sup>, [Sylvain V Costes](#)<sup>3,4</sup>

[Author information](#) [Article notes](#) [Copyright and License information](#)

PMCID: PMC6371294 PMID: [30815061](#)

### Abstract

---

Omics data sharing is crucial to the biological research community, and the last decade or two has seen a huge rise in collaborative analysis systems, databases, and knowledge bases for omics and other systems biology data. We assessed the “FAIRness” of NASA’s GeneLab Data Systems (GLDS) along with four similar kinds of systems in the research omics data domain, using 14 FAIRness metrics. The range of overall FAIRness scores was 6-12 (out of 14), average 10.1, and standard deviation 2.4. The range of Pass ratings for the metrics was 29-79%, Partial Pass 0-21%, and Fail 7-50%. The systems we evaluated performed the best in the areas of data findability and accessibility, and worst in the area of data interoperability. Reusability of metadata, in particular, was frequently not well supported. We relate our experiences implementing semantic integration of omics data from some of the assessed systems for federated querying and retrieval functions, given their shortcomings in data interoperability. Finally, we propose two new principles that Big Data system developers, in particular, should consider for maximizing data accessibility.

### Introduction

---

Over the last decade the vast increase in outputs from “omics” (genomics, transcriptomics, proteomics, and metabolomics) assays, together with decreasing assay and data storage costs, have resulted in an exponentially

increasing volume of these “big” data. The high feature-number aspect of these big data has highlighted the importance of data sharing (or reuse) functions of these systems, as data sets often have relevance for investigating a very large number of hypotheses. Thus, developers of big data systems increasingly are focusing on the needs of users to discover, annotate, share, and analyze data. Omics data sharing is especially crucial to the biological research community, and the last decade or two has seen a huge rise in collaborative analysis systems<sup>1-3</sup>, databases<sup>4-14</sup>, and knowledge bases<sup>15-17</sup> for omics and other systems biology data.

The FAIR<sup>18</sup> principles are a set of guiding principles that are being advocated as foundational in the broad areas of data findability, accessibility, interoperability, and reusability. However, the challenges involved in using current technologies and real-world infrastructures within organizational guidelines to develop data repositories and/or data analytical systems that are “FAIR” are substantial. The FAIR principles deal with aspects of data handling such as identification, representation, and attribution. Metrics to evaluate the “FAIRness” of a data system or its design are also currently being developed<sup>19</sup>. System developers should consider familiarizing themselves with these metrics, as many of the FAIR principles can have significant implications in terms of feasible design options and operations resource planning.

We assessed the “FAIRness” of NASA’s GeneLab Data Systems (GLDS), along with four systems in the research omics data domain chosen for their similarity to the GLDS, using 14 FAIRness metrics. We relate these assessments in the context of discussion of the FAIRness concepts for the design of Big Data omics systems in particular, relating some of the challenges we faced developing the GLDS as a FAIR data system. In addition to designing a data system to be FAIR, developers of Big Data systems in particular should consider how system design can maximize data usability for its users. We propose two new design principles for FAIR systems that handle Big Data like omics data from our experiences developing the GLDS. By complying with these usability principles and the FAIR principles together, systems that manage Big Data can maximize both the usability and reusability of data.

## Methods

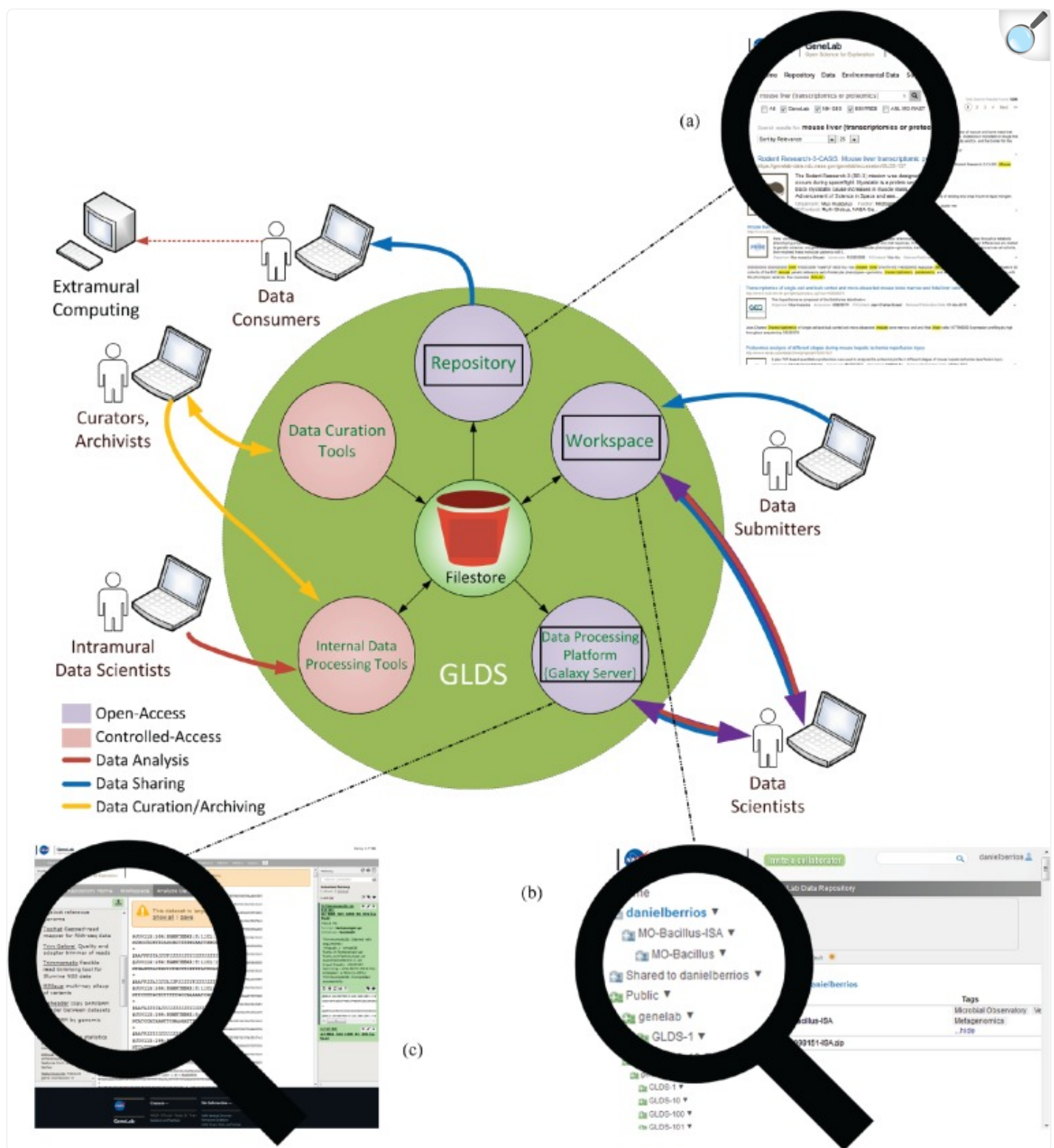
---

### NASA GeneLab

GeneLab (<http://genelab.nasa.gov>) is a NASA initiative designed to accelerate “open science” biomedical research in support of the human exploration of space and the improvement of life on earth.<sup>20</sup> GeneLab data have been used to research impacts of space on mice, humans, and plants. Phase I of GeneLab Data Systems (GLDS) project implemented an “omics” (genomics, transcriptomics, proteomics, and metabolomics) data repository for biomedical research data conducted in or relevant to space environments ([Figure 1](#)). This initial phase developed processes and systems for data submission, curation, indexing, search, and retrieval. In Phase II GeneLab implemented federated search and retrieval capabilities for data hosted by GeneLab and three other open-access, omics data systems, in order to facilitate data

integration and biological meta-investigation ([Figure 1a](#)).<sup>21</sup> (Note that these targets for search federation were selected independently and using different criteria than the data systems evaluated in this paper). Such meta-investigations are key to corroborating findings from different studies and/or kinds of bioassay data, and translating them into systems biology knowledge and, eventually, therapeutics<sup>22</sup>. In Phase III, GeneLab has begun to deploy an omics data analysis platform as part of the GLDS. Using this platform, investigators can design, execute and share in silico experimentation of omics data sets ([Figures 1b](#) and [1c](#)). At the end of Phase III, users with results of these kinds of experiments will be able to store them directly into the GLDS repository, linked to originating data and to the analysis workflows that generated them.

Figure 1.



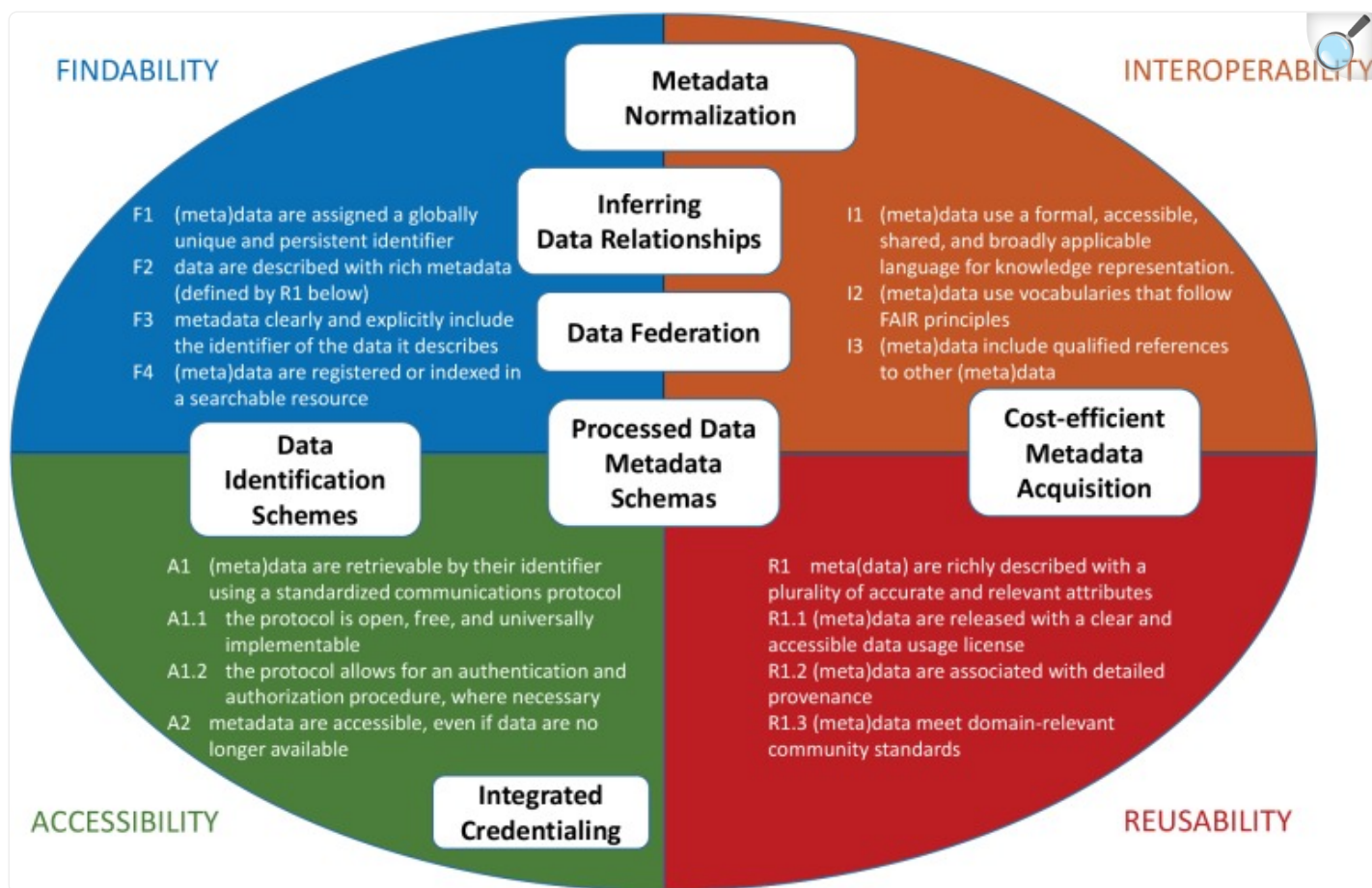
[Open in a new tab](#)

The GeneLab Data Systems (GLDS). (a) The GeneLab Data Repository. The federated search interface

shown includes not only search of GLDS repository data, but of the extramural data sources Gene Expression Omnibus (“GEO”<sup>4</sup>), PRoteomics IDentifications repository (“EBI PRIDE”<sup>5</sup>), and the Metagenomics Analysis server (“ANL MG-RAST”<sup>6</sup>). (b) The GLDS Workspace. (c) GLDS Tools for omics data analysis (part of a Galaxy server<sup>1</sup>).

Even though the inception of the GLDS predated publication of the FAIR principles, many of the concepts behind the principles were discussed extensively prior to its design. These discussions included substantial input from the National Academies of Sciences and a steering committee of outside experts in omics biological data, and had significant impacts on selection of technological approaches and system architectures, required resources for development and operations, and policies developed for operating the GLDS. In the absence of best practices from the community in certain areas of design and policy ([Figure 2](#)), we wondered how we might assess the impact on FAIRness of the design and policy choices we faced, and how the GLDS would compare in FAIRness to other open-access, Big Data systems.

Figure 2.



[Open in a new tab](#)

The FAIR principles<sup>18</sup>, and areas of design and operations of FAIR data systems (white boxes) for which best practices have yet to be defined.

## FAIRness of Omics Data

The FAIR principles were designed to guide the management of data so that it is maximally findable, accessible, interoperable, and reusable. They are not precise prescriptions for FAIR system design specifications; rather they are aspirational guidelines with some references to exemplar standards and technologies. Among the findability principles are requirements that data be assigned a persistent, global unique identifier (GUID), be described by rich metadata, and be indexed by a system for retrieving the data. Accessibility requires the use of standardized, open (available for implementation to all) communications protocols for retrieval of data by GUID. Interoperability is supported through

the use of machine-actionable metadata and data. Reusability requires systems to be compliant in the management of their data objects with the first three principles, specifically so that data and metadata can be linked and integrated across FAIR systems.

Each of the authors assessed compliance of four open-access, omics data systems and the GLDS for FAIRness using the current working draft of the FAIRness metrics developed by the GO FAIR Metrics group<sup>23</sup> (Table 1). The four systems selected were the Gene Expression Omnibus (GEO), the European Nucleotide Archive (ENA), the Metagenomics Analysis Server (MG-RAST), and Metabolights. We chose the systems for their similarity to the GLDS; all are open-access, government-operated or government-funded, research lab-developed omics data systems. As the FAIR principles themselves are only aspirational, the assessment of the systems can only be at most semi-quantitative, and they would likely be affected by further clarifications or details in the principles and/or metrics which are almost certain to come. We rated systems with Pass when we had no evidence of any failure of a test by the metric. Systems were rated as Partial Pass, if the metric had multiple steps or components, and the system passed some, but not all; or if we had evidence that some data inputs for the metric yielded a pass, while others did not. We chose Fail only if we could find no evidence the system was compliant with any part of the metric, for any inputs we tried. Individually-assessed ratings were then combined into the consensus ratings shown in [Table 1](#) through dialogue among the raters. Finally, we developed an overall FAIRness Score by assigning a score to each rating of 1 for each Pass, 0.5 for each Partial Pass, and 0 for each Fail and totaling these score for each system.



Table 1.

The FAIR principles, corresponding draft FAIRness metrics, and semi-quantitative FAIRness ratings for select omics data systems. Metrics were those developed by the GO FAIR Metrics<sup>23</sup> group.

<div> <div>● Pass</div> <div>◐ Partial pass</div> <div>○ Fail</div> </div>						
FAIR Principle	Metric	GEO <sup>4</sup>	ENA <sup>7</sup>	MG-RAST <sup>6</sup>	Metabolights <sup>8</sup>	GLDS
F1. (meta)data are assigned globally unique and persistent identifier	FM-F1A	●	●	◐	●	●
F1. (meta)data are assigned globally unique and persistent identifier	FM-F1B	●	●	○	●	●
F2. data are described with rich metadata (defined by R1 below)	FM-F2	●	●	●	●	●
F3. metadata clearly/explicitly include identifier of data it describes	FM-F3	●	●	●	●	●
F4. (meta)data are registered or indexed in a searchable resource	FM-F4	●	●	○	◐	●
A1. (meta)data are retrievable by identifier using a standardized communications protocol	N/A					
A1.1 the protocol is open, free, and universally implementable	FM-A1.1	●	●	●	●	●
A1.2 the protocol allows for an authentication and authorization procedure, where necessary	FM-A1.2	●	●	●	●	●
A2. metadata are accessible, even when data are no longer available	FM-A2	●	●	○	○	●
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.	FM-I1	○	○	○	◐	◐
I2. (meta)data use vocabularies that follow FAIR principles	FM-I2	○	○	○	◐	◐
I3. (meta)data include qualified references to other (meta)data	FM-I3	○	○	○	○	○
R1. meta(data) are richly described with a plurality of accurate and relevant attributes	N/A					
R1.1. (meta)data released with clear, accessible data usage license	FM-R1.1	●	●	●	●	●
R1.2. (meta)data are associated with detailed provenance	FM-R1.2	●	●	◐	●	●
R1.3. (meta)data meet domain-relevant community standards	FM-R1.3	●	●	○	●	●
Overall FAIRness Score		11	11	6	10.5	12

[Open in a new tab](#)

## Summary of FAIRness Assessments

There is quite some variability in FAIRness among the omics data systems we assessed using the 14 metrics. The range of Pass ratings was 29-79% of metrics, Partial Pass 0-21%, and Fail 7-50%. The range of overall FAIRness scores was



6-12, with average of 10.1 and standard deviation, 2.4. All the systems were rated best in the areas of data accessibility and reusability, slightly worse in data findability, and far worse in data interoperability.

## Findability

The first findability principle, F1 ([Table 1](#)), states that systems should identify data using globally unique and persistent identifiers; the FM-F1A metric focuses on identifier uniqueness, and FM-F1B on identifier persistence. The objective is to make data findable, but the implementation of F1 also supports data accessibility (FAIR principle A1). The qualifier “global” should be interpreted as global within the domain of conceivable use of the data. Thus, it is not sufficient to use locally (system-specific) controlled processes to manage data identifiers<sup>24</sup>, as this approach can never absolutely guarantee against overlap with other systems’ local identifier schemes. In addition, F1 specifies the persistence of issued identifiers; identification systems that employ Uniform Resource Locators (which rely on alterable domain name resolution) cannot guarantee persistence of identifiers.

Among the schemes for providing data identifiers, the most reliable over time are likely to be those that use third-party resolution (versus those that rely on in-house redirection). Third-party resolution identifiers include Digital Object Identifiers (DOIs) and those provided by [identifiers.org](#), among others<sup>25</sup>. All but one of the systems we assessed passed the metric for F1 metric by employing persistent, globally unique identifiers with third party resolution. (MG-RAST only offers such identifiers as an option during data submission, and so was assessed as Partial Pass.) In the biomedical research and clinical practice realms, publishers have now widely selected the DOI scheme for identification and citation of literature; more than 80% of 2015 articles indexed by PubMed have assigned DOIs<sup>26</sup>. As a consequence, the biomedical community is already (and continues to become more and more) familiar with using these identifiers to retrieve and cite literature. Yet to date only a few omics data repositories have similarly implemented DOIs for data publications. Consequently, actual findability and accessibility of data in these repositories is less than optimal, even though they are technically in compliance with F1, and will continue to be so, unless and until the biomedical community is trained on how to use other identifying schemes for citing data. Furthermore, metadata from these repositories are not currently indexed in DOI meta-indexes (such as the one produced by [DataCite](#)<sup>27</sup>). In order to be compliant with F1 and A1, the GLDS is implementing methods to issue DOIs for all data publications. Additionally, curated metadata records of the publications are transmitted to [DataCite](#)<sup>27</sup> for meta-indexing, increasing the chances for reuse of GLDS-hosted data.

FAIR data also need to be described “with rich metadata” in order to be findable. The draft metric to assess F2 compliance merely requires a metadata description document that specifies a format for machine-readable metadata. The definition of the term “machine-readable” varies somewhat, but it is generally agreed that it cannot be natural language. (Although F2 refers to reusability principles requiring data “have a plurality of accurate and relevant attributes” (R1), the metric for F2 does not attempt to assess compliance with principle R1 [perhaps it should]). All the systems we assessed passed this metric. Furthermore, all systems passed F3, requiring association of metadata with data

it describes. It should be noted that FM-F3 refers only to IRIs and not broadly to any other kinds of acceptable data identifiers; we assume that ultimately FM-F3 will be broadened to allow other kinds of F1-compliant data identifiers. Also, it is difficult to determine if any of the systems truly support IRIs, given that none of them currently appear to use non-ASCII character encodings in their data identifying schemes.

The FAIRness metric FM-F4 is rather ill-defined currently. It requires submission of an identifier of FAIR data and its metadata to search engines (which precisely, it does not yet specify, but some common search engines are mentioned as examples in the comments for the metric). A Pass should be given if the returned search results include links to the published data (again, it does not yet specify where in the returned results, although among both the top 10 and top 40 are mentioned in metric comments). Some of the systems we assessed failed this metric, and it seems their data records are likely not indexed by at least some search engines. For example, search of the MG-RAST data record identified by “<http://metagenomics.anl.gov/linkin.cgi?metagenome=mgm4447971.3>” or any of its tabular metadata yielded no search results referring to the repository using Google. More of the systems would fail this metric if it required data to be submitted specifically to DOI-based search systems (such as DataCite).

## Accessibility

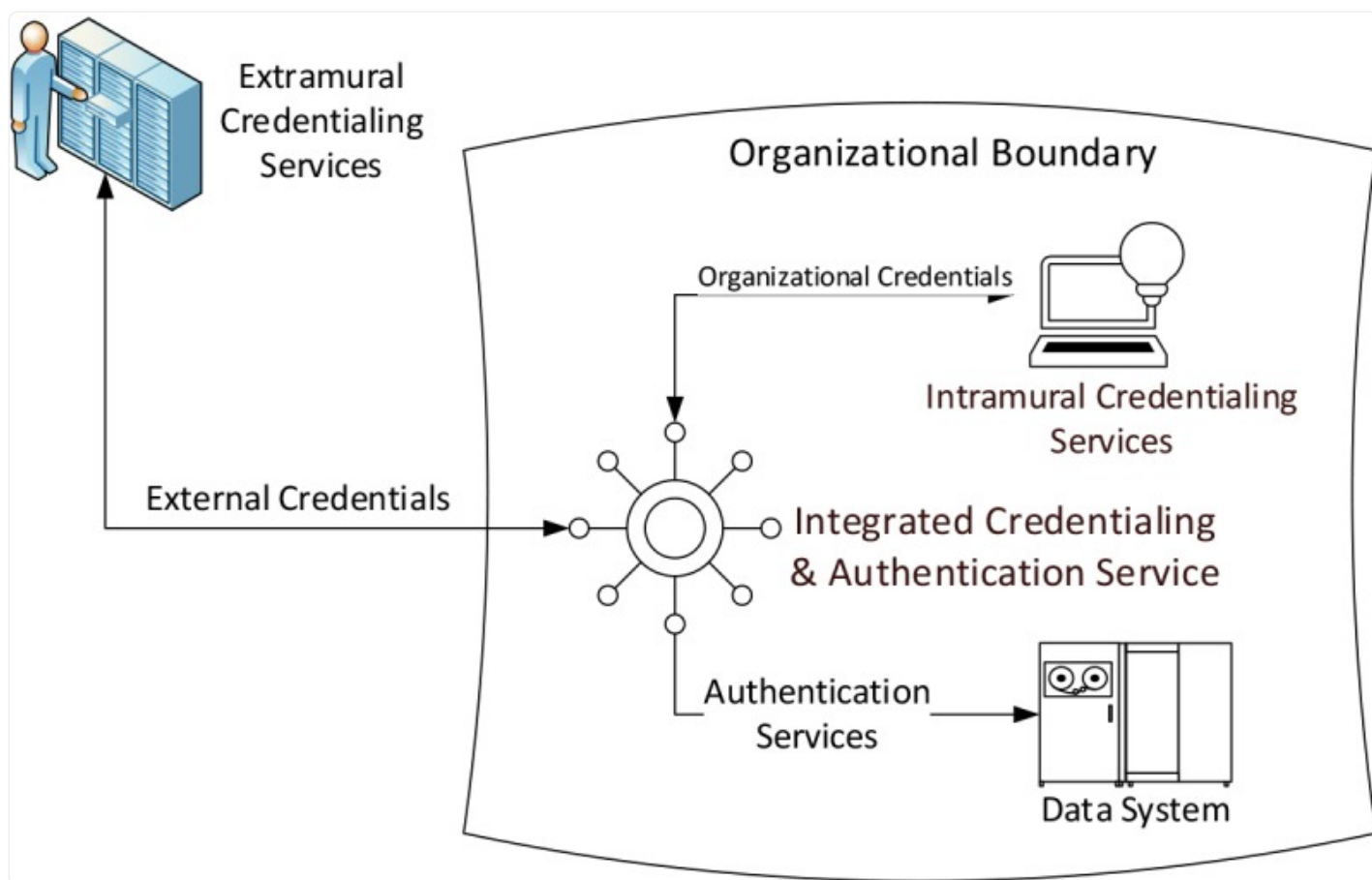
All of the systems for which we assessed compliance with FAIR principle A1.1 passed; all use open source communication protocols to access data, which are retrieved by their identifiers. FAIR Principle A1.2 stipulates that any protocol used to retrieve data “allows for an authentication and authorization procedure, where necessary.” All the systems we assessed describe which functions of the system require credentials, and how to obtain the credentials. It should be noted that A1.2 can be applied to the design of all types of systems, from those with highly restricted access controls, to the kinds of “open access” systems we assessed, that typically require minimal authentication and grant the widest possible access to all users. Furthermore, the data in question could be any kind of data, including raw instrument output, shared data analysis workflows or data products, or scientifically-relevant user comments and opinions. Even for open access systems, some systems functions such as user-workspace file management, user commenting and messaging, and user attribution of data always require authentication, because these functions must necessarily leverage user-specific information.

A specific challenge that GeneLab faced when implementing A1.2 for the GLDS was a diversity of credentialing services available to the system’s target users, with no single service common to all users. Furthermore, the goal of GeneLab is to maximize access to GLDS data across all kinds of users, but within the bounds of authentication and credentialing policies of NASA - another challenge. Aside from policy restrictions, the development of any new credentialing and authentication functions or services can require considerable resources.

A solution for organizations that are seeking to develop open-access systems is to consider investing in development of an organization-wide integrated credentialing and authentication architecture (ICAA, [Figure 3](#)), and policies for

allowing users to leverage extramural credentials, with appropriate access limitations. ICAAs and attendant policies can provide flexibility to develop and/or deploy data systems with available resources that can meet policy requirements of the organization that support extramural user access. It should be noted that using an ICAA to leverage credentialing systems developed independently from the system requires 1) technical compatibility between the authentication services of credentialing providers and the ICAA; and 2) compatibility of the credentialing policies of credentialing organizations and those of the organization developing the data system. Technical compatibility is facilitated through the use of authentication protocols and standards. Organizational credentialing policy incompatibility is more problematic to resolve. For example, it may be that the organization deploying the data system requires credentials be updated at an interval that is different from that required by one or more extramural credential provider. If these differences cannot be resolved, the only option may be to limit access for some extramural users data system functions requiring authentication.

Figure 3.



[Open in a new tab](#)

Use of an Integrated Credentialing and Authentication Architecture (ICAA) to comply with FAIR principle A1.2. The ICAA provides services and policies that use information from intra- and extramural credentialing services to authenticate users of a data system.

NASA is developing an ICAA and adopting policies that permit certain intramural data systems to leverage it for user credentials and authentication. These policies stipulates that users within the organization must use the organization's own credentialing services, while extramural users may authenticate using one from a defined set of externally-managed credentialing systems (currently including authentication by selected "social media" systems, such as Google.com credentials). At NASA (and at most U.S. federal agencies), intramural credentialing is quite a long process involving corroboration of much identifying information. The use of the ICAA, users without NASA credentials can specify a minimum of information (e.g., only a verifiable email address), which they can provide easily through a web browser, and gain almost immediate access to the GLDS. Through use of the ICAA, the GLDS is able to recognize a lower level

of confidence when authenticating extramural users, and implement appropriate access controls to some functions for these users (e.g., they may not join system administrator groups). It seems reasonable to suggest that an additional principle be considered when designing FAIR data systems that have at least some functions requiring authentication or authorization (A1.2.1 in [Table 2](#)). This principle, would stipulate that an integrated credentialing architecture, like that shown in [Figure 3](#) be leveraged, in order to offer access to data to the widest number of users possible, when and where data policies permit such access.

Table 2.

Additional principles for optimizing data accessibility

A1.2.1	Authentication protocols should support multiple credential providers
A3	Data transport should be minimized

[Open in a new tab](#)

Most of the systems we assessed for FAIR compliance to A2 have some policy to retain at least some metadata when data are removed from the system. It is frequently unclear, however, which metadata were to be retained when data records are removed from the system. Data systems should ensure that policies for retaining metadata are clearly stated, including which metadata and for what period of time.

Another factor that specifically affects accessibility to Big Data is data transport. Data to be used as inputs in Big Data systems, but which are not located in the same environment as executable code of the system, commonly must first be transported (duplicated or moved) to the environment first. This environment typically is a user-specific workspace. The transport process for Big Data can require minutes, hours or even days in some Big Data systems, and represents a significant barrier to efficient access. Designers of Big Data would benefit from a principle that stipulates the collocation within Big Data systems of data and executable code (A3 in [Table 2](#)).

## Interoperability

I1 stipulates the “use a formal, accessible, etc. language for knowledge representation. The draft metric for this principle requires that the language for data or metadata be available by URL, and have a BNF representation. The comments on the metric suggest additional required language qualities, including that the language must have semantics (“[‘vanilla’]



xml and json ..should fail”). While current knowledge representation languages such as RDF and OWL<sup>28</sup> are available to generate metadata specifications with inherent semantics, tools that leverage such languages require some sophistication in their use. All the repositories we assessed have metadata specifications that do not have inherent semantics, and thus would not yield a Pass for this metric. Metabolights and the GLDS both use the ISATab metadata specification, which can be converted to RDF, and so were rated with Partial Pass.

The metric for principle I2 has a related requirement: that data systems use open, community-developed vocabularies for data and metadata, and that such vocabularies be more than simple keywords (i.e., include semantics in some form). As the ISATab specification supported ontological lookup from such vocabularies, we again only rated Metabolights and the GLDS with Partial Pass, and the rest of the systems as FAIL.

In addition to lack of experience with knowledge representation technologies like RDF and OWL, another likely reason for lack of compliance of the systems with I1 and I2 is a general lack of available semantic resources developed by the community to generate metadata specifications and vocabularies, particularly in more recent domains like genomics, proteomics and metabolomics. We have ourselves sought such resources for use in the GLDS development, where we are seeking to provide users with omics data analysis capabilities, an omics “workbench” consisting of user-specific workspaces and data access controls, together with access to a suite of omics analysis tools (see [Figure 1b](#) and [Figure 1c](#)). Users of these tools will create data processing (analysis) workflows to transform input data into “higher-order” data products, which they will likely want to share with others and/or associate with the workflows and data inputs. One of the challenges we have observed for making these kinds of processed information FAIR include a lack of semantic resources for creating metadata specifications for such processed data and the processing workflows that generate them. While schemas like the Common Workflow Language<sup>29</sup> (CWL) specification provide a method to represent and reproduce the mechanics of data processing, they lack domain-specific semantics regarding the nature of these data transformations (e.g., what are the goals of the transformation from a biological perspective? What are the biological data concepts produced in output files? how do these relate to input data elements? etc.). Without these descriptors, systems may not adequately index these processed data and workflows for discovery and retrieval. While sophisticated users (e.g., trained bioinformaticians) who do encounter processed data and workflows may be able to understand and reuse them, others without domain-specific experience and knowledge of omics data processing may not. A challenge for the GLDS will be to help spur the creation of community resources so that it can implement rich metadata for processed omics data and workflows in a timely fashion.

The objective of the metric for principle I3 is to measure how much data in one system are linked semantically to data in other systems. It is not surprising that none of the data sources link data “out” to data in external systems in semantically qualified ways; this is a challenging objective for any data system. (We ignored links to records that were merely duplicated between systems) Manual linking of data requires subject matter expertise and can be time-consuming for systems with many data records. Automated linking requires that source and linked data both be characterized semantically so that links may be inferred. As mentioned above, semantic resources for characterizing omics data are by

and large still forthcoming.

## Reusability

R1 is referred to by the findability FAIR principle F2 (see above); at this time there is no metric specifically for R1.

R1.1 requires data systems to provide licenses for data usage (and is curiously intended to facilitate “legal interoperability” rather than reusability<sup>30</sup>). All the data repositories we assessed are governmental entities, and as such, licensing is traditionally prohibited.

R1.2 and R1.3 require that data are published together with provenance metadata, and that data and metadata conform to community-derived standards, respectively. Such standards include taxonomies, ontologies and other vocabularies for metadata, and data and file formats. The metric for R1.2 involves the demonstration that a system can provide FAIR-compliant citational and contextual provenance metadata. While all the repositories provide such metadata, some do not use FAIR-compliant schemes for these metadata. Similarly, most of the repositories assess and validate submitted data files (but few validate metadata, as discussed above) against FAIR-schemes and formats.

## Discussion

---

The systems we assessed for FAIRness clearly have shortcomings, particularly in the area of supporting data interoperability. None of the systems currently use a metadata representation language with sufficient semantics. In addition, none fully leverage FAIR vocabularies for metadata. There are perhaps several causes at work for these observations. There are often myriad vocabularies (ontologies) for metadata (each of which is a *de facto* community standard) even within a given biological sub-domain, and often much overlap in semantic content between them, creating the need for systems to be able to “normalize” input metadata across them somehow. For example, the concept of zero or very low-gravity experimental conditions appears in at least 5 different, widely-used biomedical vocabularies: SNOMED CT, MeSH, Read Codes Clinical Terms, Psychology Ontology, PLOS Thesaurus, and FAST Topical Facet, with varying language for the concept (“weightlessness” vs. “zero gravity”). True semantic interoperability and data reusability that would permit a user to find all data tagged with this concept require either that all these concept instances be individually associated with the data, or use of a system that reliably maps all the concepts from these ever-changing, ever-expanding set of ontologies. Even if such metadata normalizing mechanisms are developed, their performance and value in generating “rich” metadata is dependent on the quality of input concepts supplied either by original data submitters or biodata curators (typically employed by the system). The expertise that submitters have is critical to shaping metadata that have value for many different kinds of audiences, and that allow associated data to be found and reused. In the biomedical field, submitters to shared data repositories have relatively low motivation to go to any great lengths to provide “rich metadata” (or assist biocurators in generating them). This is likely due, at least in part, to the fact that only a small fraction of shared data appears to be cited in publication references when they are reused.<sup>31</sup>

Without their input, niche repositories and data systems, like the GLDS, with limited resources for acquiring (and maintaining) rich metadata that meet all the criteria above must constantly trade the use of resources to acquire or generate rich metadata with those that could be used for other worthy goals.

The findability and interoperability principles, when implemented in data systems, support data integration functions like multi-source data querying and retrieval. Principles F2, I1, I2, and R1.3 require the use of community-developed metadata standards, formal languages for metadata representation, FAIR vocabularies, and qualified references to other data. These principles are unfortunately not yet widely or fully implemented across biomedical data systems, making it difficult for systems to develop functions that integrate data semantically. To illustrate this, we examine our own experience developing the GLDS to provide federated search and retrieval of omics data from several data sources. We sought to import metadata records from three open-access data systems into our own metadata warehouse: GEO, MG-RAST, and the PRoteomics IDentifications repository (PRIDE)<sup>5</sup>. Each of these systems defines metadata for omics data sets differently, but none of them make these schemas available in a formal semantic language format (such as RDF or OWL). Furthermore, there are no existing maps of any of the metadata schemas to each other. As such, we resorted to manually creating semantic maps of these schemas for use in our system; queries generated via the GLDS or its RESTful search API are executed against the mapped schema concepts, and matching records are shown in a common object model representation ([Figure 4](#)). While this approach met our immediate goals of allowing users to query several data sources using our system, it required a good deal of resources and domain expertise to develop a schema mapping. Furthermore, maintaining this mapping over time as each of the data sources inevitably change will require more of our limited resources. The research community should recognize the value of federated system functions like the querying function that the GLDS offers, and urge all open access repositories and data systems to make their metadata schemas available to all according to the aforementioned FAIR principles.

Figure 4.

```

1 // 20171010164033
2 // https://genelab-data.ndc.nasa.gov/genelab/data/search?term=rodent%20liver&type=cgene,nih_geo_gse,ebi_pride,mg_rast
3
4 {
5   "took": 9,
6   "timed_out": false,
7   "_shards": {
8     "total": 5,
9     "successful": 5,
10    "failed": 0
11  },
12  "hits": {
13    "total": 65,
14    "max_score": 2.967568,
15    "hits": [
16      {
17        "_index": "glds",
18        "_type": "cgene",
19        "_id": "GLDS-137",
20        "_score": 2.967568,
21        "_source": {
22          "Authoritative Source URL": "GLDS-137/",
23          "Flight Program": "International Space Station (ISS)",
24          "Mission": {
25            "End Date": "11-May-2016",
26            "Start Date": "08-Apr-2016",
27            "Name": "SpaceX_8"
28          },
29          "Material Type": "Live",
30          "Project Identifier": "RR-3",
31          "Accession": "GLDS-137",
32          "Study Identifier": "RR3 Liver",
33          "Study Protocol Name": "sample collection      nucleic acid extraction      library construction      nucleic acid
sequencing      sequence analysis      data transformation      protein extraction      labeling      mass spectrometry      data
transformation",
34          "Study Assay Technology Type": "nucleotide sequencing      mass spectrometry      nucleotide sequencing",
35          "Acknowledgments": "",
36          "Study Assay Technology Platform": "Illumina      LTQ Orbitrap Velos (Thermo Scientific)      Illumina",
37          "Study Person": {
38            "Last Name": "Globus      GeneLab      Smith      Cramer",
39            "Middle Initials": "",
40            "First Name": "Ruth      NASA      Rosamund      Marty"
41          },
42          "Study Protocol Type": "sample collection      nucleic acid extraction      library construction      nucleic acid
sequencing      sequence analysis      data transformation      protein extraction      labeling      mass spectrometry      data
transformation",
43          "Space Program": "NASA",
44          "Study Title": "Rodent Research-3-CASIS: Mouse liver transcriptomic proteomic and epigenomic data",
45          "Study Factor Type": "Weightlessness, eatme",
46          "Study Public Release Date": "1503903600",

```

[Open in a new tab](#)

Federated Search of Omics Data in the GLDS. The RESTful search API result shows 65 records from the four data sources searched (GLDS, GEO, PRIDE, and MG-RAST). Each records metadata is mapped using a single schema for display in the GLDS (see [Figure 1a](#)).

## References

---

1. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, et al. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.* 2005;15(10):1451–5. doi: 10.1101/gr.4086505. [[DOI](#)] [[PMC free article](#)] [[PubMed](#)] [[Google Scholar](#)]
2. Arkin AP, Stevens RL, Cottingham RW, Maslov S, Henry CS, Dehal P, et al. 2016. The DOE Systems Biology Knowledgebase (KBase). *bioRxiv*. [[Google Scholar](#)]
3. Goff SA, Vaughn M, McKay S, Lyons E, Stapleton AE, Gessler D, et al. The iPlant Collaborative: Cyberinfrastructure for Plant Biology. *Front Plant Sci.* 2011;2:34. doi: 10.3389/fpls.2011.00034. [[DOI](#)] [[PMC free article](#)] [[PubMed](#)] [[Google Scholar](#)]
4. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 2002;30(1):207–10. doi: 10.1093/nar/30.1.207. [[DOI](#)] [[PMC free article](#)] [[PubMed](#)] [[Google Scholar](#)]
5. Jones P, Cote RG, Martens L, Quinn AF, Taylor CF, Derache W, et al. PRIDE: a public repository of protein and peptide identifications for the proteomics community. *Nucleic Acids Res.* 2006;34((Database issue)):D659–63. doi: 10.1093/nar/gkj138. [[DOI](#)] [[PMC free article](#)] [[PubMed](#)] [[Google Scholar](#)]
6. Meyer F, Paarmann D, D’Souza M, Olson R, Glass EM, Kubal M, et al. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics.* 2008;9(1):386. doi: 10.1186/1471-2105-9-386. [[DOI](#)] [[PMC free article](#)] [[PubMed](#)] [[Google Scholar](#)]
7. Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tarraga A, Cheng Y, et al. The European Nucleotide Archive. *Nucleic Acids Res.* 2011;39((Database issue)):D28–31. doi: 10.1093/nar/gkq967. [[DOI](#)] [[PMC free article](#)] [[PubMed](#)] [[Google Scholar](#)]
8. Haug K, Salek RM, Conesa P, Hastings J, de Matos P, Rijnbeek M, et al. MetaboLights--an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res.* 2013;41((Database issue)):D781–6. doi: 10.1093/nar/gks1004. [[DOI](#)] [[PMC free article](#)] [[PubMed](#)] [[Google Scholar](#)]
9. Motenko H, Neuhauser SB, O’Keefe M, Richardson JE. MouseMine: a new data warehouse for MGI. *Mamm Genome.* 2015;26(7-8):325–30. doi: 10.1007/s00335-015-9573-z. [[DOI](#)] [[PMC free article](#)] [[PubMed](#)] [[Google Scholar](#)]
10. Stein L, Sternberg P, Durbin R, Thierry-Mieg J, Spieth J. WormBase: network access to the genome and



- biology of *Caenorhabditis elegans*. *Nucleic Acids Res.* 2001;29(1):82–6. doi: 10.1093/nar/29.1.82. [[DOI](#)] [[PMC free article](#)] [[PubMed](#)] [[Google Scholar](#)]
11. Arend D, Junker A, Scholz U, Schuler D, Wylie J, Lange M. 2016 2016. PGP repository: a plant phenomics and genomics data publication infrastructure. *Database (Oxford)* [[DOI](#)] [[PMC free article](#)] [[PubMed](#)] [[Google Scholar](#)]
12. Leinonen R, Sugawara H, Shumway M. The sequence read archive. *Nucleic Acids Res.* 2011;39((Database issue)):D19–21. doi: 10.1093/nar/gkq1019. [[DOI](#)] [[PMC free article](#)] [[PubMed](#)] [[Google Scholar](#)]
13. Huala E, Dickerman AW, Garcia-Hernandez M, Weems D, Reiser L, LaFond F, et al. The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res.* 2001;29(1):102–5. doi: 10.1093/nar/29.1.102. [[DOI](#)] [[PMC free article](#)] [[PubMed](#)] [[Google Scholar](#)]
14. FlyBase: the *Drosophila* database. *Nucleic Acids Res.* 1996;24(1):53–6. doi: 10.1093/nar/24.1.53. [[DOI](#)] [[PMC free article](#)] [[PubMed](#)] [[Google Scholar](#)]
15. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, et al. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.* 2005;33((Database issue)):D428–32. doi: 10.1093/nar/gki072. [[DOI](#)] [[PMC free article](#)] [[PubMed](#)] [[Google Scholar](#)]
16. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 1999;27(1):29–34. doi: 10.1093/nar/27.1.29. [[DOI](#)] [[PMC free article](#)] [[PubMed](#)] [[Google Scholar](#)]
17. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* 2000;25(1):25–9. doi: 10.1038/75556. [[DOI](#)] [[PMC free article](#)] [[PubMed](#)] [[Google Scholar](#)]
18. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data.* 2016;3:160018. doi: 10.1038/sdata.2016.18. [[DOI](#)] [[PMC free article](#)] [[PubMed](#)] [[Google Scholar](#)]
19. Wilkinson MD, Sansone S-A, Schultes E, Doorn P, Bonino da Silva, Santos LO, Dumontier M. 2017. A design framework and exemplar metrics for FAIRness. *bioRxiv*. [[DOI](#)] [[PMC free article](#)] [[PubMed](#)] [[Google Scholar](#)]
20. Berrios D, Thompson T, Fogle H, Rask J, Coughlan J. San Francisco: Nov 14-18, GeneLab: NASA's Open Access, Collaborative Platform for Systems Biology and Space Medicine. Poster presented at: 2015 AMIA

Annual Symposium; p. CA2015. [[Google Scholar](#) ]

21. Berrios D, Welch J, Fogle H, Skidmore M, Marcu O. Chicago: GeneLab: NASA's GeneLab: Phase I Results and Plans. Poster presented at: 2016 AMIA Annual Symposium; p. IL2016. [[Google Scholar](#) ]

22. Beheshti A, Cekanaviciute E, Smith DJ, Costes SV. Global transcriptomic analysis suggests carbon dioxide as an environmental stressor in spaceflight: a systems biology GeneLab case study. *Sci Rep*. 2018;8(1):4191. doi: 10.1038/s41598-018-22613-1. [[DOI](#) ] [[PMC free article](#)] [[PubMed](#)] [[Google Scholar](#) ]

23. FAIR Metrics ALL: GO FAIR Metrics Group. 2018. [updated January 11, 2018. Available from: [[DOI](#) ]

24. McMurry JA, Juty N, Blomberg N, Burdett T, Conlin T, Conte N, et al. Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. *PLoS Biol*. 2017;15(6):e2001414. doi: 10.1371/journal.pbio.2001414. [[DOI](#) ] [[PMC free article](#)] [[PubMed](#)] [[Google Scholar](#) ]

25. Starr J, Castro E, Crosas M, Dumontier M, Downs RR, Duerr R, et al. 2015. Achieving human and machine accessibility of cited data in scholarly publications. *PeerJ Comput Sci*; p. 1. [[DOI](#) ] [[PMC free article](#)] [[PubMed](#)] [[Google Scholar](#) ]

26. Boudry C, Chartron G. Availability of digital object identifiers in publications archived by PubMed. *Scientometrics*. 2017;110(3):1453–69. [[Google Scholar](#) ]

27. Neumann J, Brase J. DataCite and DOI names for research data. *J Comput Aided Mol Des*. 2014;28(10):1035–41. doi: 10.1007/s10822-014-9776-5. [[DOI](#) ] [[PubMed](#)] [[Google Scholar](#) ]

28. Patel-Schneider PF, Hayes P, Horrocks I. OWL Web Ontology Language: Semantics and Abstract Syntax. 2004;W3C [[Google Scholar](#) ]

29. Peter A, Michael R. C, Nebojša T, Brad C, John C, Michael H, et al. Common Workflow Language. v1:02016. [[Google Scholar](#) ]

30. R1.1: (Meta)data are released with a clear and accessible data usage license: GO FAIR Initiative. [Available from: <https://www.go-fair.org/fair-principles/r1-1-metadata-released-clear-accessible-data-usage-license/>]

31. Kratz JE, Strasser C. Researcher perspectives on publication and peer review of data. *PLoS One*. 2015;10(2):e0117619. doi: 10.1371/journal.pone.0117619. [[DOI](#) ] [[PMC free article](#)] [[PubMed](#)] [[Google Scholar](#) ]

---

Articles from AMIA Annual Symposium Proceedings are provided here courtesy of **American Medical Informatics Association**